Given a text, find the number of authors based on writing style changes. Using word patterns, discern an exclusive personal style of each author. Presence of multiple styles indicates presence of multiple authors.

Problem Description

Dataset characteristics

• StackExchange forum posts in English.
• 2546 (training) / 1272 (validation) documents.
• 1-5 authors.
• Mean number of tokens: 1570.

Strategy

Preliminary steps
1. Tokenize a document into windows of paragraph length (Fig. 1).
2. Extract 50 Most Frequent Words (MFW) from each window.
3. Compare distance between windows (Matusita) and create distance matrix (see Fig. 2, 4).

Threshold Based Clustering Algorithm (TBC)
• Cluster by selecting windows with smallest distances iteratively (when forming a cluster, the closest members are included first).
• Discriminates later potential members using cluster thresholds (Fig. 5).

Window Merge Clustering Algorithm (WMC)
• Iteratively combine most similar windows to generate a new set of windows.
• Recalculate distance matrix from these new windows, for the next iteration. As a result, the clusters formed are hierarchical (Fig. 6).

Evaluation

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Acc.</th>
<th>OCI</th>
<th>Rank</th>
<th>Acc.</th>
<th>OCI</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBC</td>
<td>0.66</td>
<td>0.83</td>
<td>0.42</td>
<td>0.65</td>
<td>0.82</td>
<td>0.42</td>
</tr>
<tr>
<td>WMC</td>
<td>0.62</td>
<td>0.91</td>
<td>0.35</td>
<td>0.63</td>
<td>0.88</td>
<td>0.37</td>
</tr>
<tr>
<td>Combined Min</td>
<td>0.65</td>
<td>0.92</td>
<td>0.36</td>
<td>0.66</td>
<td>0.9</td>
<td>0.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Acc.</th>
<th>OCI</th>
<th>Rank</th>
<th>Acc.</th>
<th>OCI</th>
<th>Rank</th>
<th>Acc.</th>
<th>OCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBC</td>
<td>0.83</td>
<td>0.87</td>
<td>0.57</td>
<td>0.83</td>
<td>0.85</td>
<td>0.49</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>WBC</td>
<td>0.72</td>
<td>0.93</td>
<td>0.74</td>
<td>0.72</td>
<td>0.9</td>
<td>0.42</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Combined Min</td>
<td>0.70</td>
<td>0.93</td>
<td>0.39</td>
<td>0.72</td>
<td>0.91</td>
<td>0.41</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Conclusion

• TBC performed the best out of three approaches (improvements were observed with duplicated sentences information).
• Winner of the PAN CLEF 2019 SCD challenge.
• Demonstrated that it is possible to identify the number of authors from a relatively short piece of text without any prior training corpus per author.

Contact: Sukanya Nath, Computer Science Department
sukanya.nath@unine.ch

Web: http://www.unine.ch
University of Neuchâtel